SEMI-PARAMETRIC ESTIMATION IN THE BLOCK PARALLEL MISSING DATA GRAPHICAL MODEL

by

Jinchang Fan

A thesis submitted to The Johns Hopkins University in conformity with the requirements for the degree of Master of Science.

Baltimore, Maryland

December, 2019

Abstract

This paper developed and implemented estimating equation based estimation methods for a Missing Not At Random (MNAR) model, in particular estimator based on Influence Functions (IFs) in the Block Parallel Missing Data graphical model. Experiments show that estimators based on Influence Function (IF) make a significant improvement over more common Inverse-Probability Weighted (IPW) estimators, especially in models contain more than two variables. Though performance of Efficient Influence Function(EIF) estimator is not so good and due to the complexity it is currently limited to model with only two variables, we believe a generalized version for models with more variables will still be valuable.

Primary Reader and Advisor: Ilya Shpitser

Acknowledgments

I thank Ilya Shpitser for guiding me into causal inference, which changed my way of understanding the world. For his humor, his deep thoughts, his patience and his amicability. I thank Lin Liu for his earlier work on this topic and selflessly sharing with me. I thank my lab mates in Ilya's group: Rohit Bhattacharya, Razieh Nabi, Numair Sani, Dan Malinsky and Jaron Lee for discussing my questions and creating laugh in the group.

Dedication

I am appreciated to be born and live in the peaceful part of this world, so I don't need to worry about livelihood and safety. I thank my family, especially my mother, for providing me with the opportunity to be well educated and encouraging me to be myself. I thank my beloved partner Yakun Deng for supporting me since we met. I thank my friends, Yuxuan Zhang, Luqiao Xu, Paiheng Xu and Yashil Sukurdeep for encouraging me in my hard time and bringing warmness and laugh to me.

Contents

Abstract ii

Acknowledgments iii

List of Tables vii

List of Figures viii

1 Introduction 1

2 The Block Parallel Missing Data Graphical Model 3

2.1 Graphical Model Methods for Missing Data 3

2.2 Directed Acyclic Graph 4

2.3 Missing Data Graphical Models 5

2.4 Identification and the Block Parallel Missing Data graphical model 6

3 Influence Functions 10

4 The Efficient Influence Function 13

CONTENTS

5 Experimental Results **16**

5.1 Two Variables Simulation 17

5.2 Three Variables Simulation 18

5.3 Six Variables Simulation 19

5.4 Summary 20

6 Conclusion **22**

A Derivation of nonparametric Influence Function **23**

B Derivation of Efficient Influence Function in two-variables model **27**

Bibliography **37**

Vita **39**

List of Tables

5.1 Result of estimators in two-variable model 17
5.2 Result of estimators in three-variable model 19
5.3 Result of estimators in six-variable model 20

List of Figures

2.1 the Block Parallel Missing Data graphical model with k variables 7
2.2 the Block Parallel Missing Data graphical model with two variables 9

5.1 Total error and variance of IFs in two-variable model 18
5.2 Total error and variance of IFs in three-variable model 19
5.3 Total error and variance of IFs in six-variable model 20

Chapter 1

Introduction

In practice, the data generating process is not always fully observed. Missing data is ubiquitous in practical data analysis problems. This problem is especially serious when variables are censored systematically, as it would lead to biased estimation if observations with missing entries are simply ignored or discarded. Standard assumptions about the mechanism of missingness include assuming missingness is completely independent of observed and unobserved data, which is known as Missing Completely At Random (MCAR) model, or assuming it is independent of unobserved data given observed data, which is known as Missing At Random (MAR) model. But these models are insufficient when missingness depends on values that are themselves unobserved, known as Missing Not At Random (MNAR), which is a frequent situation in real world data. Karthika Mohan and Judea Pearl have described an MNAR model which

provides a simple condition for the target distribution to be identified [1,2] and for a special case of two-variable model, Lin Liu and Jamie Robins have provided the corresponding estimator characterized by Influence Functions (IFs). In this paper, the IF based estimator is extended to model with arbitrary number of variables, and implemented on synthetic data to testify its correctness and compare the efficiency with other estimators such as Inverse-Probability Weighted (IPW) estimator.

The rest of this paper is organized as follows: Chapter 2 will discuss Block Parallel Missing Data graphical model of missing data and estimators based on it. Chapter 3 will provide the IF estimator of this model, and Chapter 4 will reach to the most efficient one that attains the semi-parametric efficiency bounds. Simulation result on models with different number of variables is shown in Chapter 5, and Chapter 6 is conclusion on the whole work.

Chapter 2

The Block Parallel Missing Data Graphical Model

2.1 Graphical Model Methods for Missing Data

Causal inference is concerned with expressing causal effects of an intervention operation as functionals of the observed data distribution.It often helps to view missing data problems through a causal lense by considering this problem as computing distributions of variables, had they been counterfactually observed, from actual data, where variables are possibly censored. A Directed Acyclic Graph (DAG) which represents constraints in a causal model can be

used to represent constraints in the mechanism of missingness as well, with these constraints potentially allowing identification of the full data distribution.

2.2 Directed Acyclic Graph

A DAG is a graph $\mathcal{G}$ with vertices in set $\mathbf{V}$ represent random variables connected by directed edges in set $\mathbf{E}$, where no directed cycle is allowed. To be clear about the statement some terms are introduced here: in a graph $\mathcal{G}$ with a set of vertices $\mathbf{V}$, parents of vertex V is $\mathrm{pa}_{\mathcal{G}}(V) \equiv \{U \in \mathbf{V} | U \to V\}$, children of vertex V is $\mathrm{ch}_{\mathcal{G}}(V) \equiv \{U \in \mathbf{V} | V \to U\}$, descendents of vertex V is $\mathrm{de}_{\mathcal{G}}(V) \equiv \{U \in \mathbf{V} | V \to \cdots \to U\}$, ancestors of vertex V is $\mathrm{an}_{\mathcal{G}}(V) \equiv \{U \in \mathbf{V} | U \to \cdots \to V\}$, non-descendants of vertex $\mathbf{V}$ is $\mathrm{nd}_{\mathcal{G}}(V) \equiv \mathbf{V} \setminus \mathrm{de}_{\mathcal{G}}(V)$.

A statistical model of a DAG $\mathcal{G}$ is set of distributions $p(\mathbf{V})$ on random variables such that $p(\mathbf{V}) = \prod_{V \in \mathbf{V}} p\left(V | \mathrm{pa}_{\mathcal{G}}(V)\right)$. Causal model of a DAG is similar set of distributions on counterfactual variables. Given random variable $Y \in \mathbf{V}$ and $A \in \mathbf{V} \setminus \{Y\}$, $p(Y(a))$ represents the distribution of counterfactual random variables $Y(a)$, which is the distribution of Y in a hypothetical situation where variable A were intervened and set to be value a [3].

2.3 Missing Data Graphical Models

For missing data problems with k variables, there are two sets of variables: $\mathbf{L}^{(1)} = \{L_1^{(1)}, L_2^{(1)}, \cdots, L_k^{(1)}\}$ and $\mathbf{R} = \{R_1, R_2, \cdots, R_k\}$, representing target random variables and indicators of their missingness respectively. Each random variables in set $\mathbf{L}^{(1)}$ is potentially missing with corresponding observed proxy variable L_i, defined as $L_i \equiv L_i^{(1)}$ if $R_i = 1$, and $L_i \equiv$ "?" if $R_i = 0$, where $R_i \in \mathbf{R}$ is corresponding indicator random variable. In missing data problems, the goal is to estimate the target distribution $p(\mathbf{L}^{(1)})$, based on observed distribution $P(\mathbf{L}, \mathbf{R})$. It is equivalent to estimating distribution $p(\mathbf{L})$ in a counterfactual situation where all indicators $R \in \mathbf{R}$ were set to be one. The corresponding missing data graph is a DAG that

- every observed variables $L_i \in \mathbf{L}$ has and only has two parents $L_i^{(1)}$ and R_i
- missingness indicators $\mathbf{R}$ have to outgoing edge toward counterfactual variables $\mathbf{L}^{(1)}$
- observed variables $\mathbf{L}$ have no outgoing edge

2.4 Identification and the Block Parallel Missing Data graphical model

The distribution $p(\mathbf{L}^{(1)})$ is said to be identified if could be written as a function of observed data distribution $P(\mathbf{L}, \mathbf{R})$. One condition of the distribution to be identified is that it can be decomposed into admissible factorization, which is an ordered factorization, or a sum of such factorizations such that every factor $p_i = p(L_i^{(1)}|L_j^{(1)})$ satisfies $L_i^{(1)} \perp\!\!\!\perp (R_j, R_i)|L_j^{(1)}$. Then each factor p_i can be identified as observed distribution $p(L_i|L_j, R_j = 1, R_i = 1)$. This condition is a sufficient condition but not complete, in the sense that it fails to identify $p\left(\mathbf{L}^{(1)}\right)$ in certain identifiable models. Mohan, Tian, and Pearl [4] gave a more general, necessary and sufficient condition for identifying $p(\mathbf{L}^{(1)})$ in a DAG $\mathcal{G}$ which is there is with no edge between missingness indicators and no edge between counterfactual target variables and missingness indicator of itself: $R_i \to R_j \notin \mathbf{E}, \forall i, j \in \{1, 2, \cdots, k\}; L_i^{(1)} \to R_i \notin \mathbf{E}, \forall i \in \{1, 2, \cdots, k\}$. In this model shown in Figure 2.1, by the Markov property of DAG each R_i is independent of $L_i^{(1)}$ and other missingness indicators $\mathbf{R}_{-i}$ given all the other $L_j^{(1)}$: $R_i \perp\!\!\!\perp L_i^{(1)}, \mathbf{R}_{-i}|\mathbf{L}_{-i}^{(1)}, \forall i \in \{1, 2, \cdots, k\}$. Here $\mathbf{L}_{-i}$ means all $L^{(1)}$ except $L_i^{(1)}$: $\{L_1^{(1)}, L_2^{(1)}, \cdots, L_{i-1}^{(1)}, L_{i+1}^{(1)}, \cdots, L_k^{(1)}\}$, and similar for $\mathbf{R}_{-i}$. Without further notation $\Pr(R_i = 1| \cdot)$ and $\Pr(\cdot |R_i = 1)$ will be short-written as $\Pr(R_i| \cdot)$ and $\Pr(\cdot |R_i)$ respectively. The same notation will be used for the remaining part of

this paper. Then $p(\mathbf{L}^{(1)})$ could be identified as

$$p(\mathbf{L}^{(1)}) = \frac{\Pr(\mathbf{R}=1)\Pr(\mathbf{L}^{(1)},\mathbf{R})}{\prod_i \Pr\left(R_i|\mathbf{L}^{(1)}_{-i},\mathbf{R}_{-i}\right)} = \frac{(\prod \mathbf{R})\Pr(\mathbf{L},\mathbf{R})}{\prod_i \Pr\left(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i}\right)} \tag{2.1}$$

And for arbitrary function $h(\mathbf{L}^{(1)})$, expectation $\mathbb{E}(h(\mathbf{L}^{(1)}))$ could be estimated by taking the following (inverse weighted) empirical average:

$$\begin{aligned}\mathbb{E}(h(\mathbf{L}^{(1)})) &= \mathbb{E}\left\{h\left(\mathbf{L}^{(1)}\right)\frac{(\prod \mathbf{R})\Pr(\mathbf{L},\mathbf{R})}{\prod_i \Pr\left(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i}\right)}\right\} \\ &= \frac{1}{N}\sum_j^N \frac{(\prod \mathbf{R}_j)h\left(\mathbf{L}^{(1)}_j\right)}{\prod_i \Pr\left(R_{i,j}|\mathbf{L}_{-i,j},\mathbf{R}_{-i,j}\right)}\end{aligned} \tag{2.2}$$

It is called Inverse-Probability Weighted (IPW) estimation which is a consistent estimator of the function $h(\mathbf{L}^{(1)})$ if all the model $\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i}=1;\theta_i), i \in \{1,2,\cdots,k\}$ are correct.

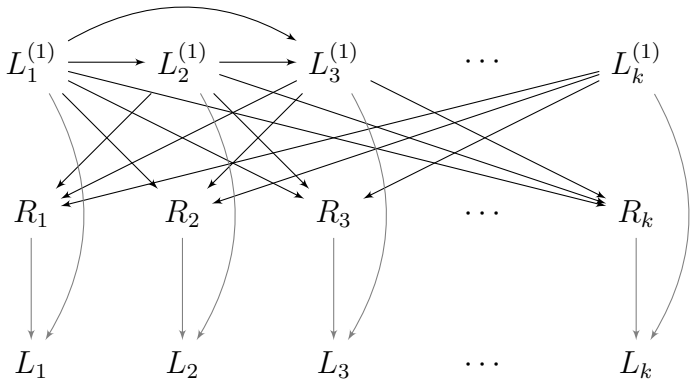


Figure 2.1: the Block Parallel Missing Data graphical model with k variables

A simple example of Block Parallel Missing Data graphical model with two

variables is shown in Figure 2.2, where the following conditional independences exist: $R_1 \perp\!\!\!\perp (L_1^{(1)}, R_2)|L_2^{(1)}$, and $R_2 \perp\!\!\!\perp (L_2^{(1)}, R_1)|L_1^{(1)}$. By these conditional independences $P(L_1^{(1)}, L_2^{(1)})$ is shown to be identified as

$$\begin{aligned}
\Pr(L_1^{(1)}, L_2^{(1)}) &= \Pr(L_1^{(1)}, L_2^{(1)}) \frac{\Pr(R_1, R_2|L_1^{(1)}, L_2^{(1)})}{\Pr(R_1, R_2|L_1^{(1)}, L_2^{(1)})} \\
&= \frac{\Pr(R_1, R_2) \Pr(L_1^{(1)}, L_2^{(1)}|R_1, R_2)}{\Pr(R_1, R_2|L_1^{(1)}, L_2^{(1)})} \\
&= \frac{\Pr(R_1, R_2) \Pr(L_1^{(1)}, L_2^{(1)}|R_1, R_2)}{\Pr\left(R_1|R_2, L_2^{(1)}\right) \Pr\left(R_2|R_1, L_1^{(1)}\right)} \\
&= \frac{\Pr(R_1, R_2) \Pr(L_1, L_2|R_1, R_2)}{\Pr\left(R_1|R_2, L_2\right) \Pr\left(R_2|R_1, L_1\right)}
\end{aligned} \tag{2.3}$$

And for arbitrary function $h(L_1, L_2)$, expectation $\mathbb{E}(h(L_1^{(1)}, L_2^{(1)}))$ could be estimated by taking the empirical average:

$$\begin{aligned}
\mathbb{E}(h(L_1^{(1)}, L_2^{(1)})) &= \mathbb{E}\left\{ h\left(L_1, L_2\right) \frac{R_1 R_2 \Pr(L_1, L_2|R_1, R_2)}{\Pr\left(R_1|R_2, L_2\right) \Pr\left(R_2|R_1, L_1\right)} \right\} \\
&= \frac{1}{N} \sum_j^N \frac{R_{1j} R_{2j} h\left(L_{1j}, L_{2j}\right)}{\Pr\left(R_{1j}|R_{2j}, L_{2j}\right) \Pr\left(R_{2j}|R_{1j}, L_{1j}\right)}
\end{aligned} \tag{2.4}$$

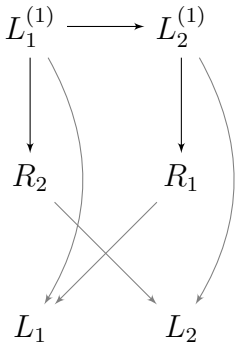


Figure 2.2: the Block Parallel Missing Data graphical model with two variables

Chapter 3

Influence Functions

Let $\mathbf{Z} = \{\mathbf{L}, \mathbf{R}\}$, $Z_1, \cdots, Z_n$ be i.i.d. samples from a general class of probability densities $p(Z;\theta)$ parameterized by $\theta^T = (\beta^T, \eta^T)$, where $\beta \in \mathbb{R}^q$ denotes the set of target parameters, and η denotes a possible infinite dimensional set of nuisance parameters. This type of model is called semi-parametric model, as it has both a parametric component β and a non-parametric one η. The goal of statistical inference in semi-parametric models is to find "the best" estimator of β in this model, denoted by $\hat{\beta}$. Regular Asymptotically Linear (RAL) estimators are considered with the form

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi\left(Z_i\right) + o_p(1),$$

where $o_p(1)$ is a term that converges in probability to zero as n goes to infinity and $\phi(.) \in \mathbb{R}^q$ is an Influence Function (IF) with mean zero and finite variance. There is a bijective correspondence between RAL estimators and IFs as RAL estimators are consistent and asymptotically normal with the variance of the estimator given by its IF:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \phi\phi^T\right).$$

For the interested parameter $\beta = h(\mathbf{L^{(1)}})$ where $h(\cdot)$ is an arbitrary function, from equation (2.1) it can be identified as

$$\beta = h(\mathbf{L^{(1)}}) = \frac{\mathrm{P}(\mathbf{R})h(\mathbf{L}^{(1)}, \mathbf{R})}{\prod_i \Pr\left(R_i|\mathbf{L}_{-i}^{(1)}, \mathbf{R}_{-i}\right)} = \frac{(\prod \mathbf{R})h(\mathbf{L}, \mathbf{R})}{\prod_i \Pr\left(R_i|\mathbf{L}_{-i}, \mathbf{R}_{-i}\right)} \tag{3.1}$$

Denote full data influence function for β as $IF(\mathbf{L^{(1)}}, \mathbf{R})$ and the score functions as $S_\theta(\mathbf{L}, \mathbf{R})$. Then the relationship between IF and S_θ is [5]

$$\mathbb{E}\left[IF \cdot S_\theta\right] = \frac{\partial \mathbb{E}\left[\frac{(\prod \mathbf{R})h(\mathbf{L})}{\prod_i \Pr(R_i|\mathbf{L}_{-i}, \mathbf{R}_{-i})}\right]}{\partial \theta} \tag{3.2}$$

Denote full data influence function for β as $IF(\mathbf{L^{(1)}}, \mathbf{R})$ and the score functions as $S_\theta(\mathbf{L}, \mathbf{R})$. Then the relationship between IF and S_θ is [5]

$$\mathbb{E}\left[IF \cdot S_\theta\right] = \frac{\partial \mathbb{E}\left[\frac{(\prod \mathbf{R})h(\mathbf{L})}{\prod_i \Pr(R_i|\mathbf{L}_{-i}, \mathbf{R}_{-i})}\right]}{\partial \theta} \tag{3.3}$$

Without any further restriction, the nonparametric influence function IF^* can be derived: (see Appendix A)

$$U(\beta) = \left[\frac{(\prod \mathbf{R}) h(\mathbf{L})}{\prod_i \Pr\left(R_i | \mathbf{L}_{-i}, \mathbf{R}_{-i}\right)} - \mathbb{E}\left\{ \frac{(\prod \mathbf{R}) h(\mathbf{L})}{\prod_i \Pr\left(R_i | \mathbf{L}_{-i}, \mathbf{R}_{-i}\right)} \right\} \right] + \sum_{i=1}^{k} \left[\left(1 - \frac{R_i}{\Pr\left(R_i | \mathbf{L}_{-i}, \mathbf{R}_{-i}\right)} \right) \prod \mathbf{R}_{-i} \mathbb{E}\left\{ \frac{h\left(\mathbf{L}\right)}{\prod_{j \neq i} \Pr\left(R_j | \mathbf{L}_{-j}, \mathbf{R}_{-j}\right)} | \mathbf{L}_{-i}, \mathbf{R}_{-i} \right\} \right] \tag{3.4}$$

Since $\mathbb{E}[U(\beta)] = 0$, the derived estimator for β is

$$\mathbb{E}\left\{ \frac{(\prod \mathbf{R}) h(\mathbf{L})}{\prod_i \Pr\left(R_i | \mathbf{L}_{-i}, \mathbf{R}_{-i}\right)} + \sum_{i=1}^{k} \left[\left(1 - \frac{R_i}{\Pr\left(R_i | \mathbf{L}_{-i}, \mathbf{R}_{-i}\right)} \right) \prod \mathbf{R}_{-i} \mathbb{E}\left\{ \frac{h\left(\mathbf{L}\right)}{\prod_{j \neq i} \Pr\left(R_j | \mathbf{L}_{-j}, \mathbf{R}_{-j}\right)} | \mathbf{L}_{-i}, \mathbf{R}_{-i} \right\} \right] \right\} \tag{3.5}$$

Chapter 4

The Efficient Influence Function

In the previous section, IF^* is derived without any restriction. But actually Block Parallel Missing Data Graph Model is not saturated, as conditional independence between R imply a constraint on the observed data distribution here [6]. Thus there exists a group of IFs that satisfies equation (3.3), and the IF with smallest variance will provide the most efficient estimation. In fact, IFs provide a geometric view of the behavior of RAL estimators. Consider a Hilbert space $\mathcal{H}$ of all mean zero q-dimensional functions, equipped with an inner product of two arbitrary elements h_1, h_2 of the Hilbert space defining as $\mathbb{E}\left[h_1^T h_2\right]$. Define a parametric submodel to be a subset of densities in the semi-parametric model parameterized by $\theta_\gamma^T = \left(\beta^T, \gamma^T\right)$, where $\gamma^T \in \mathbb{R}^r$ such that the subset contains the density $p(Z;\theta_0)$ in the semi-parametric model evaluated at the true parameter values θ_0. The nuisance tangent space Λ in the semi-parametric

model is defined to be the linear subspace generated by the nuisance score vector $S_\eta(z,\theta_0)$: $\Lambda = \{B^{q\times r}S_\eta(Z,\theta_0)$ for all $B^{q\times r}\}$ where $S_\eta(z,\theta_0) = \frac{\partial \log p_Z(z,\theta)}{\partial \eta}\Big|_{\theta=\theta_0}^{r\times 1}$ and $B^{q\times r}$ is all $q \times r$ matrices. The space Λ is important because it is known that all IFs lie in the orthogonal complement $\Lambda^\perp$ of Λ with respect to $\mathcal{H}$ [5]. Out of all IFs in $\Lambda^\perp$ there exists a unique IF which is orthogonal to observed data tangent space so it has smallest variance matrix. It is called Efficient Influence Function (EIF).

For a two-variable model the EIF is given by Lin Liu as follows (see Appendix B).

$$EIF^{obs} = IF^* - \mathbb{E}\left\{IF^* \cdot g\left(R_1, L_1R_1, R_2, L_2R_2\right)^T\right\} \cdot \left[\mathbb{E}\left\{g\left(R_1, L_1R_1, R_2, L_2R_2\right) \cdot g\left(R_1, L_1R_1, R_2, L_2R_2\right)^T\right\}\right]^{-1} \cdot g\left(R_1, L_1R_1, R_2, L_2R_2\right) \tag{4.1}$$

where $g(R_1, R_1L_1, R_2, R_2L_2)$ is the vector that lies in the orthogonal tangent space with an arbitrary constant scalar g_0:

$$
\begin{aligned}
& g\left(R_1, R_1 L_1, R_2, R_2 L_2\right) = \\
& R_1 R_2 \left\{ \frac{\left(1-\pi_{R_2}\left(L_1\right)\right)\left(1-\mathbb{E}\left[\pi_{R_1}\left(L_2\right)|L_1\right]\right)}{\pi_{R_2}\left(L_1\right)\mathbb{E}\left[\pi_{R_1}\left(L_2\right)|L_1\right]} + \frac{\left(1-\pi_{R_1}\left(L_2\right)\right)\left(1-\mathbb{E}\left[\pi_{R_2}\left(L_1\right)|L_2\right]\right)}{\pi_{R_1}\left(L_2\right)\mathbb{E}\left[\pi_{R_2}\left(L_1\right)|L_2\right]} \right. \\
& \qquad \left. - \frac{\left(1-\pi_{R_1}\left(L_2\right)\right)\left(1-\pi_{R_2}\left(L_1\right)\right)}{\pi_{R_1}\left(L_2\right)\pi_{R_2}\left(L_1\right)} \right\} \cdot g_0 \\
& - (1-R_1)(R_2)\left\{\mathbb{E}\left[\pi_{R_2}\left(L_1\right)|L_2\right]\right\}^{-1}\left\{1-\mathbb{E}\left[\pi_{R_2}\left(L_1\right)|L_2\right]\right\} \cdot g_0 \\
& - R_1(1-R_2)\left\{\mathbb{E}\left[\pi_{R_1}\left(L_2\right)|L_1\right]\right\}^{-1}\left\{1-\mathbb{E}\left[\pi_{R_1}\left(L_2\right)|L_1\right]\right\} \cdot g_0 \\
& + (1-R_1)\left(1-R_2\right) g_0
\end{aligned}
\tag{4.2}
$$

Chapter 5

Experimental Results

In this chapter estimators based on IFs are evaluated on synthetic data, generated by parametric bootstrap with 100 replicates. In the following sections there are simulations with two, three, and six variables respectively, evaluated by total error, standard error, width of 95% confidence interval with bootstrap, and median of variance of IF (for IPW estimator, it is the variance of equation (3.1)). Here $\Pr(R_i|\mathbf{L}_{-i}, \mathbf{R}_{-i})$ is estimated by logistic regression which is the correctly specified parametric model in the data generating process, and other expectations are estimated by nonparametric kernel regression. To verify the generality of estimation method in all simulations variables are generated dependent to each other and target parameter is a non-linear function of variables.

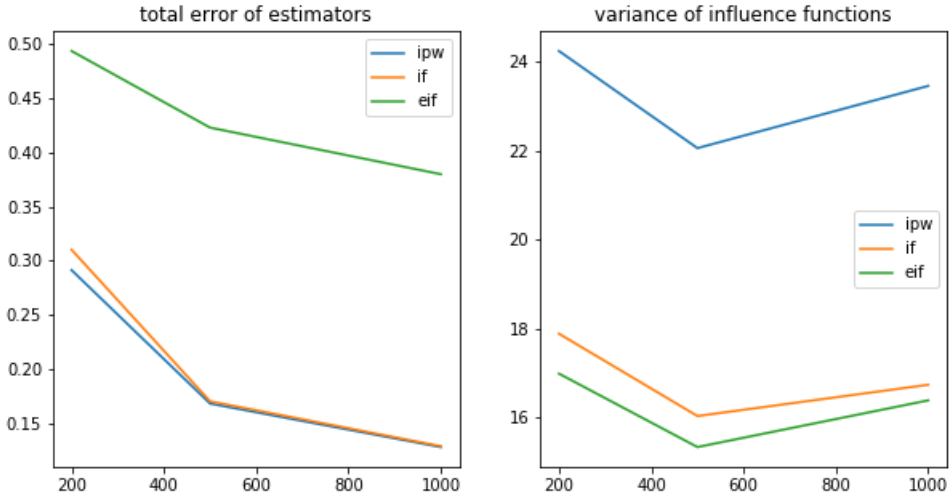
5.1 Two Variables Simulation

In Simulation 1 a two-variable model is implemented, with around 33% of the data are fully observed, 50% of the data are observed with one variable and 10% are fully censored. The true parameter that we are interested in is 3.

Table 5.1: Result of estimators in two-variable model

Estimator	Data size	Estimation	T.E.	S.E.	Variance of IF	Width of C.I. with bootstrap
IPW	200	3.02	0.28	0.28	25.87	1.18
	500	3.01	0.19	0.19	25.11	0.73
	1000	2.99	0.12	0.12	24.54	0.41
IF	200	3.02	0.28	0.28	15.51	1.16
	500	3.0	0.19	0.19	14.83	0.7
	1000	2.99	0.12	0.12	14.48	0.36
EIF	200	3.12	0.39	0.37	15.08	1.44
	500	3.2	0.32	0.25	14.71	0.97
	1000	3.18	0.24	0.15	14.36	0.54

Figure 5.1: Total error and variance of IFs in two-variable model



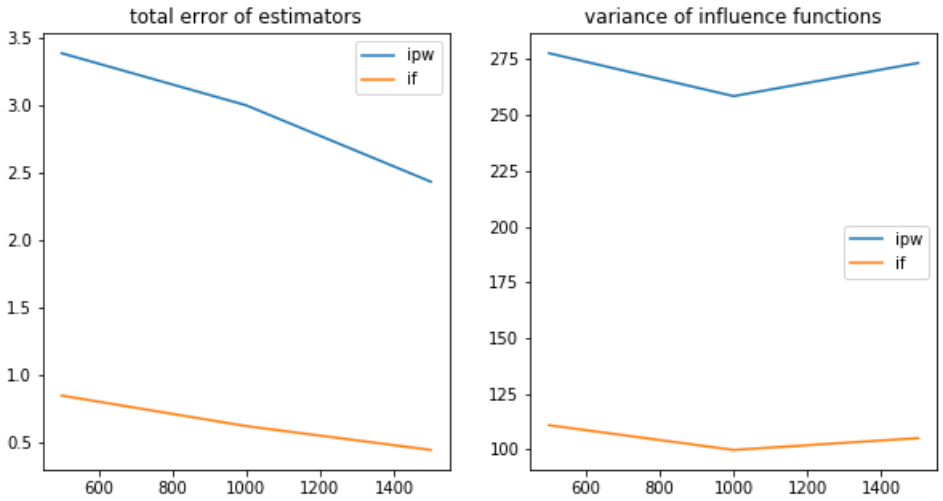
5.2 Three Variables Simulation

In Simulation 2, a three-variable model is implemented, with around 36% of the data are fully observed and 42% of the data are observed with two variable. The true parameter that we are interested in is 10.

Table 5.2: Result of estimators in three-variable model

Estimator	Data size	Estimation	T.E.	S.E.	Variance of IF	Width of C.I. with bootstrap
IPW	500	11.68	3.38	2.94	277.72	11.49
	1000	11.24	3.0	2.73	258.47	10.64
	1500	11.14	2.43	2.15	273.34	8.88
IF	500	9.78	0.84	0.81	110.94	2.86
	1000	9.7	0.62	0.54	99.86	2.22
	1500	9.81	0.44	0.4	105.16	1.67

Figure 5.2: Total error and variance of IFs in three-variable model



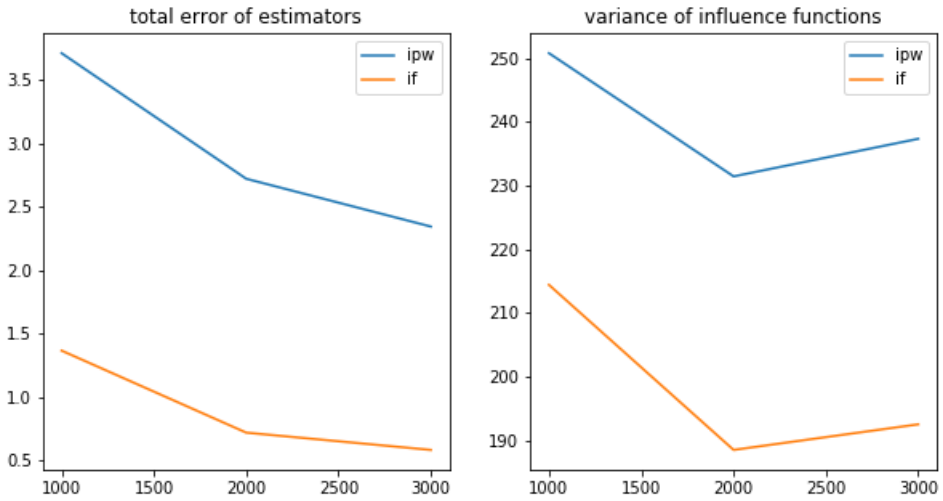
5.3 Six Variables Simulation

In Simulation 3, a six-variable model is implemented, with around 25% of the data are fully observed and 40% of the data are observed with five variable. The true parameter that we are interested in is 8.

Table 5.3: Result of estimators in six-variable model

Estimator	Data size	Estimation	T.E.	S.E.	Variance of IF	Width of C.I. with bootstrap
IPW	1000	9.42	3.71	3.43	250.82	12.42
	2000	9.14	2.72	2.47	231.46	9.08
	3000	8.9	2.34	2.17	237.37	7.99
IF	1000	7.7	1.37	1.33	214.42	3.8
	2000	7.73	0.72	0.67	188.49	2.54
	3000	7.76	0.59	0.53	192.51	1.38

Figure 5.3: Total error and variance of IFs in six-variable model



5.4 Summary

It shows that in terms of accuracy the performance of IPW estimator and IF estimator are quite similar in two-variable model, but the advantage of IF estimator shows more clear when number of variables increases. That is reasonable if we notice that the fully observing rate is going down when more

variables come into model, as IPW estimator only uses information from data that are fully observed, while IF estimator takes extra information from data that only one variable is censored.

In a two-variables model the total error of estimator based on EIF is unexpectedly larger than estimator based on IF and IPW. It may come from the noise when fitting a regression model $\mathbb{E}[\Pr(R_2 = 1|L_1)|L_2]$, where the dependent variable $\Pr(R_2 = 1|L_1)$ itself is fitted from a regression model as well. Variance of EIF does be smaller than variance of If and IPW, but the improvement of EIF over IF is not so significant compared with the improvement of IF over IPW. Though its performance is not ideal in this model, EIF estimator may still be valuable in models with more variables, considering the fact its taking use of information from observations with all possible situations of missingness, whereas IPW and IF estimators only use information from part of whole observations.

Chapter 6

Conclusion

In this paper estimators based on Influence Function are deduced for the Block Parallel Missing Data graphical model with an arbitrary number of variables, as well as the estimator based on the Efficient Influence Function with two variables. In comparison with ordinary Inverse-Probability Weighted estimators, synthetic data experiment has verified the significant advantage of estimators based on Influence Function, in terms of accuracy and variance of estimators, especially in model with more variables. Followup work would entail deriving the EIF for the model with more than two variables.

Appendix A

Derivation of nonparametric Influence Function

According to equation (3.3):

$$\mathbb{E}\left[IF \cdot S_\theta\right] = \frac{\partial \mathbb{E}\left[\frac{(\prod \mathbf{R})h(\mathbf{L})}{\prod_i \Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i};\theta_i)}\right]}{\partial \theta}$$

Now look into the derivative of $\mathbb{E}\left[\frac{(\prod \mathbf{R})h(\mathbf{L})}{\prod_i \Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i};\theta_i)}\right]$ corresponding to $\Pr(R_1|\mathbf{L}_{-1},\mathbf{R}_{-1};\theta_1)$:

APPENDIX A. DERIVATION OF NONPARAMETRIC INFLUENCE FUNCTION

$$
\begin{aligned}
&\frac{\partial}{\partial\theta_1}\mathbb{E}\left\{\frac{(\prod\mathbf{R})h(\mathbf{L})}{\prod_i \Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}\right\}\\
=&\frac{\partial}{\partial\theta_1}\int\frac{(\prod\mathbf{R})h(\mathbf{L})}{\prod_i \Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}d\mathbb{P}(\mathbf{R},\mathbf{L})\\
=&\int-\frac{(\prod\mathbf{R})h(\mathbf{L})}{\Pr^2(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\prod_{i\neq1}\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}\frac{\partial\Pr(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}{\partial\theta_1}\cdot d\mathbb{P}(\mathbf{R},\mathbf{L})\\
=&\int-\frac{(\prod\mathbf{R})h(\mathbf{L})}{\Pr^2(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\prod_{i\neq1}\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}\frac{\partial\mathbb{E}(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}{\partial\theta_1}d\mathbb{P}(\mathbf{R},\mathbf{L})\\
=&\int-\frac{(\prod\mathbf{R})h(\mathbf{L})}{\Pr^2(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\prod_{i\neq1}\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}\cdot\left\{\int R_1\frac{\partial f(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1})}{\partial\theta_1}dR_1\right\}d\mathbb{P}(\mathbf{R},\mathbf{L})\\
=&\int-\frac{(\prod\mathbf{R})h(\mathbf{L})}{\Pr^2(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\prod_{i\neq1}\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}\\
&\cdot\left\{\int R_1\frac{f(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}{f(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}\frac{\partial f(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}{\partial\theta_1}dR_1\right\}d\mathbb{P}(\mathbf{R},\mathbf{L})\\
=&\int-\frac{(\prod\mathbf{R})h(\mathbf{L})}{\Pr^2(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\prod_{i\neq1}\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}\\
&\cdot\left\{\int R_1\frac{\partial\log f(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}{\partial\theta_1}f(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\,dR_1\right\}d\mathbb{P}(\mathbf{R},\mathbf{L})\\
=&\int-\frac{(\prod\mathbf{R})h(\mathbf{L})}{\Pr^2(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\prod_{i\neq1}\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}\\
&\cdot\mathbb{E}\left\{R_1\cdot S(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\,|\mathbf{R}_{-1},\mathbf{L}_{-1}\right\}d\mathbb{P}(\mathbf{R},\mathbf{L})\\
=&\int-\frac{(\prod\mathbf{R})h(\mathbf{L})}{\prod_i\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}\\
&\cdot\mathbb{E}\left\{\frac{R_1-\Pr(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}{\Pr(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}\cdot S(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\,|\mathbf{R}_{-1},\mathbf{L}_{-1}\right\}d\mathbb{P}(\mathbf{R},\mathbf{L})\\
=&\int\frac{(\prod\mathbf{R})h(\mathbf{L})}{\prod_i\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}\\
&\cdot\mathbb{E}\left\{\left(1-\frac{R_1}{\Pr(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}\right)\cdot S(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\,|\mathbf{R}_{-1},\mathbf{L}_{-1}\right\}d\mathbb{P}(\mathbf{R},\mathbf{L})
\end{aligned}
$$

(continued from previous page)

$$
\begin{aligned}
=&\int \frac{(\prod_{i\neq 1}\mathbf{R})h(\mathbf{L})}{\prod_{i\neq 1}\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}\mathbb{E}\left\{\left(1-\frac{R_1}{\Pr(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}\right)\cdot S(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\,|\mathbf{R}_{-1},\mathbf{L}_{-1}\right\}\\
&\cdot f(L_1|R_1,\mathbf{R}_{-1},\mathbf{L}_{-1})\,f(\mathbf{R}_{-1},\mathbf{L}_{-1})\,(\prod_{i\neq 1}d\mathbf{R})(\prod d\mathbf{L})\\
=&\mathbb{E}\left[\mathbb{E}\left\{\frac{(\prod_{i\neq 1}\mathbf{R})h(\mathbf{L})}{\prod_{i\neq 1}\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}\cdot\mathbb{E}\left\{\left(1-\frac{R_1}{\Pr(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}\right)\right.\right.\right.\\
&\left.\left.\left.\cdot S(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\,|R_1=1,\mathbf{R}_{-1},\mathbf{L}_{-1}\right.\right.\right.\\
=&\mathbb{E}\left[\mathbb{E}\left\{\frac{(\prod_{i\neq 1}\mathbf{R})h(\mathbf{L})}{\prod_{i\neq 1}\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}|R_1,\mathbf{R}_{-1},\mathbf{L}_{-1}\right\}\right.\\
&\left.\cdot\mathbb{E}\left\{\left(1-\frac{R_1}{\Pr(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}\right)\cdot S(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\,|\mathbf{R}_{-1},\mathbf{L}_{-1}\right\}\right]\\
=&\mathbb{E}\left[\mathbb{E}\left\{\mathbb{E}\left\{\frac{(\prod_{i\neq 1}\mathbf{R})h(\mathbf{L})}{\prod_{i\neq 1}\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}|R_1,\mathbf{R}_{-1},\mathbf{L}_{-1}\right\}\right.\right.\\
&\left.\left.\cdot\left(1-\frac{R_1}{\Pr(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}\right)\cdot S(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\,|\mathbf{R}_{-1},\mathbf{L}_{-1}\right\}\right]\\
=&\mathbb{E}\left\{\left(1-\frac{R_1}{\Pr(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)}\right)(\prod_{i\neq 1}\mathbf{R})\mathbb{E}\left\{\frac{(\prod_{i\neq 1}\mathbf{R})h(\mathbf{L})}{\prod_{i\neq 1}\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}|R_1,\mathbf{R}_{-1},\mathbf{L}_{-1}\right\}\right.\\
&\left.\cdot S(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1};\theta_1)\right.
\end{aligned}
$$

Thus the part of the IF* corresponding to $\Pr(R_1|\mathbf{L}_{-1},\mathbf{R}_{-1};\theta_1)$ is

$$\left(1-\frac{R_1}{\Pr(R_1|\mathbf{R}_{-1},\mathbf{L}_{-1})}\right)(\prod_{i\neq 1}\mathbf{R})\mathbb{E}\left\{\frac{(\prod_{i\neq 1}\mathbf{R})h(\mathbf{L})}{\prod_{i\neq 1}\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i})}|\mathbf{R}_{-1},\mathbf{L}_{-1}\right\}\qquad\textbf{(A.1)}$$

And similarly the part corresponding to $\Pr(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i};\theta_i)$ is

APPENDIX A. DERIVATION OF NONPARAMETRIC INFLUENCE FUNCTION

$$\left(1-\frac{R_i}{\Pr\left(R_i|\mathbf{R}_{-i},\mathbf{L}_{-i}\right)}\right)\left(\prod_{j\neq i}\mathbf{R}\right)\mathbb{E}\left\{\frac{\left(\prod_{j\neq i}\mathbf{R}\right)h\left(\mathbf{L}\right)}{\prod_{j\neq i}\Pr\left(R_j|\mathbf{L}_{-j},\mathbf{R}_{-j}\right)}|\mathbf{R}_{-j},\mathbf{L}_{-j}\right\} \tag{A.2}$$

Thus the nonparametric Influence Function IF^* is

$$\begin{aligned} IF^* = & \left[\frac{\left(\prod\mathbf{R}\right)h(\mathbf{L})}{\prod_i\Pr\left(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i}\right)}-\mathbb{E}\left\{\frac{\left(\prod\mathbf{R}\right)h(\mathbf{L})}{\prod_i\Pr\left(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i}\right)}\right\}\right] \\ & +\sum_{i=1}^{k}\left[\left(1-\frac{R_i}{\Pr\left(R_i|\mathbf{L}_{-i},\mathbf{R}_{-i}\right)}\right)\prod\mathbf{R}_{-i}\mathbb{E}\left\{\frac{h\left(\mathbf{L}\right)}{\prod_{j\neq i}\Pr\left(R_j|\mathbf{L}_{-j},\mathbf{R}_{-j}\right)}|\mathbf{L}_{-i},\mathbf{R}_{-i}\right\}\right] \end{aligned} \tag{A.3}$$

Appendix B

Derivation of Efficient Influence Function in two-variables model

To find all the other influence functions constrained by the conditional independencies induced from the missing data graph, we can try to find the space of functions Γ orthogonal to the observed data tangent space formed by the score functions Γ^{obs}. The complete data tangent space of the score function of equation (**??**) can be written as

$$\begin{aligned}\Gamma^{full} = \{&S_{L_1,L_2}(L_1, L_2) + S_{L_2}(R_1|L_2) + S_{L_1}(R_2|L_1); \\ &\mathbb{E}[S_{L_1,L_2}(L_1, L_2)] = 0, \mathbb{E}[S_{L_2}(R_1|L_2)] = 0, \mathbb{E}[S_{L_1}(R_2|L_1)] = 0\}.\end{aligned} \tag{B.1}$$

And the corresponding observed data tangent space of the score function is just to regress the complete data score function with respect to the observed data

APPENDIX B. DERIVATION OF EFFICIENT INFLUENCE FUNCTION IN TWO-VARIABLES MODEL

$(R_1, L_1R_1, R_2, L_2R_2)$:

$$\begin{aligned}\Gamma^{obs} = \{&\mathbb{E}[S_{L_1,L_2}(L_1, L_2) + S_{L_2}(R_1|L_2) + S_{L_1}(R_2|L_1)|R_1, R_2, L_1R_1, L_2R_2]; \\ &\mathbb{E}[S_{L_1,L_2}(L_1, L_2)] = 0, \mathbb{E}[S_{L_2}(R_1|L_2)] = 0, \mathbb{E}[S_{L_1}(R_2|L_1] = 0\}\end{aligned} \tag{B.2}$$

Hence we can further write down the functions in the observed score tangent space. For convenience denote $\Pr(R_1 = 1|L_2, R_2 = 1)$ as $\pi_{R_1}(L_2)$, and similarly $\Pr(R_2 = 1|L_1, R_1 = 1)$ as $\pi_{R_2}(L_1)$.

APPENDIX B. DERIVATION OF EFFICIENT INFLUENCE FUNCTION IN TWO-VARIABLES MODEL

$$
\begin{aligned}
& S_{obs}(R_1, R_1L_1, R_2, R_2L_2) \\
=& \mathbb{E}[S_{L_1,L_2}(L_1, L_2) + S_{R_1,L_2}(R_1|L_2) + S_{R_2,L_1}(R_2|L_1)|R_1, R_2, L_1R_1, L_2R_2] \\
=& R_1R_2[S_{L_1,L_2}(L_1, L_2) + S_{R_1,L_2}(R_1|L_2) + S_{R_2,L_1}(R_2|L_1)] \\
& + R_1(1-R_2)[\mathbb{E}[S_{L_1,L_2}(L_1, L_2)|R_1 = 1, L_1, R_2 = 0] + \mathbb{E}[S_{R_1,L_2}(R_1|L_2)|R_1 = 1, L_1, R_2 = 0] \\
& \qquad + S_{R_2,L_1}(R_2|L_1)] \\
& + (1-R_1)R_2[\mathbb{E}[S_{L_1,L_2}(L_1, L_2)|R_1 = 0, R_2 = 1, L_2] + S_{R_1,L_2}(R_1|L_2) \\
& \qquad + \mathbb{E}[S_{R_2,L_1}(R_2|L_1)|R_1 = 0, R_2 = 1, L_2] \\
& + (1-R_1)(1-R_2)[\mathbb{E}[S_{L_1,L_2}(L_1, L_2)|R_1 = 0, R_2 = 0] + \mathbb{E}[S_{R_1,L_2}(R_1|L_2)|R_1 = 0, R_2 = 0] \\
& \qquad + \mathbb{E}[S_{R_2,L_1}(R_2|L_1)|R_1 = 0, R_2 = 0] \\
=& R_1R_2\{S_{L_1,L_2}(L_1, L_2) + [R_1 - \pi_{R_1}(L_2)]h_{L_2}(L_2) + [R_2 - \pi_{R_2}(L_1)]h_{L_1}(L_1)\} \\
& + (1-R_1)R_2\{\mathbb{E}[S_{L_1,L_2}(L_1, L_2)|R_1 = 0, R_2 = 1, L_2] + [R_1 - \pi_{R_1}(L_2)]h_{L_2}(L_2) \\
& \qquad + \mathbb{E}[[R_2 - \pi_{R_2}(L_1)]h_{L_1}(L_1)|R_1 = 0, R_2 = 1, L_2\} \\
& + R_1(1-R_2)\{\mathbb{E}[S_{L_1,L_2}(L_1, L_2)|R_1 = 1, L_1, R_2 = 0] \\
& \qquad + \mathbb{E}[[R_1 - \pi_{R_1}(L_2)]h_{L_2}(L_2)|R_1 = 1, L_1, R_2 = 0] + [R_2 - \pi_{R_2}(L_1)]h_{L_1}(L_1)\} \\
& + (1-R_1)(1-R_2)\{\mathbb{E}[S_{L_1,L_2}(L_1, L_2)|R_1 = 0, R_2 = 0] \\
& \qquad + \mathbb{E}[[R_1 - \pi_{R_1}(L_2)]h_{L_2}(L_2)|R_1 = 0, R_2 = 0] \\
& \qquad + \mathbb{E}[[R_2 - \pi_{R_2}(L_1)]h_{L_1}(L_1)|R_1 = 0, R_2 = 0]\}
\end{aligned}
\tag{B.3}
$$

APPENDIX B. DERIVATION OF EFFICIENT INFLUENCE FUNCTION IN TWO-VARIABLES MODEL

Lemma 1. *The observed data orthogonal complement* $\Gamma^{,obs,1}$ *against the observed data tangent space* Γ^{obs} *is equivalent to the observed data orthogonal complement* $\Gamma^{,obs,2}$ *against the full data tangent space* Γ^{full}.

Proof.

$$\begin{aligned}
&\mathbb{E}\{g(R_1L_1, R_1, R_2L_2, R_2)S_{obs}(R_1L_1, R_1, R_2L_2, R_2)\} \\
=&\mathbb{E}\{g(R_1L_1, R_1, R_2L_2, R_2)\mathbb{E}[S_{full}(L_1, R_1, L_2, R_2)|R_1L_1, R_1, R_2L_2, R_2]\} \\
=&\mathbb{E}\{\mathbb{E}[g(R_1L_1, R_1, R_2L_2, R_2)S_{full}(L_1, R_1, L_2, R_2)|R_1L_1, R_1, R_2L_2, R_2]\} \\
=&\mathbb{E}\{g(R_1L_1, R_1, R_2L_2, R_2)S_{full}(L_1, R_1, L_2, R_2)\}.
\end{aligned}$$

where the first equality follows from Lemma 2 □

Lemma 2. *The observed data tangent space* Γ^{obs} *is the projection of the full data tangent space* Γ^{full} *onto the observed data, i.e. with an abuse of notation* $\Gamma^{obs} = \mathbb{E}[\Gamma^{full}|O]$*, denoting the observed data vector as* O.

Proof. Let an arbitrary $S^{full}(L, R) \in \Gamma^{full}$, then by definition $S^{full}(L, R) = \frac{\partial log\mathcal{L}_\theta^{full}}{\partial\theta}$ and $\mathbb{E}(S^{full}) = 0$. As a consequence, we also have $\mathbb{E}\{\mathbb{E}(S^{full}|O)\} = \mathbb{E}(S^{full}) = 0$.

APPENDIX B. DERIVATION OF EFFICIENT INFLUENCE FUNCTION IN TWO-VARIABLES MODEL

And we can write down the observed data likelihood function as

$$
\begin{aligned}
\mathcal{L}_\theta^{obs} &= \int_{\{R=0\}} \mathcal{L}_\theta^{full}(L,R)dL \\
S^{obs} &= \frac{\partial}{\partial\theta} log\mathcal{L}_\theta^{obs} = \frac{\partial}{\partial\theta} log \int_{\{R=0\}} \mathcal{L}_\theta^{full}(L,R)dL \\
&= \frac{\frac{\partial}{\partial\theta}\int_{\{R=0\}} \mathcal{L}_\theta^{full}(L,R)dL}{\int_{\{R=0\}} \mathcal{L}_\theta^{full}(L,R)dL} = \frac{\int_{\{R=0\}} \frac{\partial}{\partial\theta}\mathcal{L}_\theta^{full}(L,R)dL}{\int_{\{R=0\}} \mathcal{L}_\theta^{full}(L,R)dL} \\
&= \frac{\int_{\{R=0\}} S^{full}\mathcal{L}_\theta^{full}(L,R)dL}{\int_{\{R=0\}} \mathcal{L}_\theta^{full}(L,R)dL} = \int_{\{R=0\}} S^{full}\mathcal{L}_\theta(L,R|O)dL \\
&= \mathbb{E}(S^{full}|O)
\end{aligned}
$$

□

To derive Γ, it is easier to derive the Γ for each term of equation B.3 and then find the intersection among the spaces. The orthogonal complement against

the observed score tangent space due to $\Pr(R_1 = 1|L_2, R_2)$ is

$$
\begin{aligned}
0 =& \mathbb{E}[R_1 R_2 (1 - \Pr(R_1 = 1|L_2, R_2)) h_{L_2}(L_2) g_{L_1,L_2}(L_1, L_2)] \\
&+ \mathbb{E}[R_1(1 - R_2)\mathbb{E}[(1 - \Pr(R_1 = 1|L_2, R_2)) h_{L_2}(L_2)|R_1 = 1, L_1, R_2 = 0] \cdot g_{L_1}(L_1)] \\
&- \mathbb{E}[(1 - R_1) R_2 \Pr(R_1 = 1|L_2, R_2) h_{L_2}(L_2) g_{L_2}(L_2)] \\
&- \mathbb{E}[(1 - R_1)(1 - R_2)\mathbb{E}[\Pr(R_1 = 1|L_2, R_2) h_{L_2}(L_2)|R_1 = 0, R_2 = 0] g_0] \\
=& \mathbb{E}[R_1 R_2 (1 - \Pr(R_1 = 1|L_2, R_2)) h_{L_2}(L_2) g_{L_1,L_2}(L_1, L_2)] \\
&+ \mathbb{E}[R_1(1 - R_2)(1 - \Pr(R_1 = 1|L_2, R_2)) h_{L_2}(L_2) g_{L_1}(L_1)] \\
&- \mathbb{E}[(1 - R_1) R_2 \Pr(R_1 = 1|L_2, R_2) h_{L_2}(L_2) g_{L_2}(L_2)] \\
&- \mathbb{E}[(1 - R_1)(1 - R_2) \Pr(R_1 = 1|L_2, R_2) h_{L_2}(L_2) g_0] \\
=& \mathbb{E}[\pi_{R_1}(L_2)\pi_{R_2}(L_1)(1 - \pi_{R_1}(L_2)) h_{L_2}(L_2) g_{L_1,L_2}(L_1, L_2)] \\
&+ \mathbb{E}[\pi_{R_1}(L_2)(1 - \pi_{R_2}(L_1))(1 - \pi_{R_1}(L_2)) h_{L_2}(L_2) g_{L_1}(L_1)] \\
&- \mathbb{E}[(1 - \pi_{R_1}(L_2))\pi_{R_2}(L_1)\pi_{R_1}(L_2) h_{L_2}(L_2) g_{L_2}(L_2)] \\
&- \mathbb{E}[(1 - \pi_{R_1}(L_2))(1 - \pi_{R_2}(L_1))\pi_{R_1}(L_2) h_{L_2}(L_2) g_0]
\end{aligned}
$$

where the second equality follows from the duality between observed tangent space and full tangent space as in Lemma 2, and the third equality holds by using the Tower's law for iterative expectation. Thus we can obtain the first constraint:

$$\mathbb{E}[\pi_{R_2}(L_1)\{g_{L_1,L_2}(L_1, L_2) - g_{L_2}(L_2)\} + (1 - \pi_{R_2}(L_1))\{g_{L_1}(L_1) - g_0\}|R_2, L_2] = 0 \quad \text{(B.4)}$$

APPENDIX B. DERIVATION OF EFFICIENT INFLUENCE FUNCTION IN TWO-VARIABLES MODEL

Similarly we can obtain another constraint by symmetry:

$$\mathbb{E}[\pi_{R_1}(L_2)\{g_{L_1,L_2}(L_1,L_2)-g_{L_1}(L_1)\}+(1-\pi_{R_1}(L_2))\{g_{L_2}(L_2)-g_0\}|R_1,L_1]=0 \quad \text{(B.5)}$$

We also obtain the third constraint by finding the orthogonal complement with respect to $\mathbb{E}[S_{L_1,L_2}(L_1,L_2)|R_1,L_1R_1,R_2,L_2R_2]$:

$$\begin{aligned}
&\mathbb{E}[S_{L_1,L_2}(L_1,L_2)\{\pi_{R_1}(L_2)\pi_{R_2}(L_1)g_{L_1,L_2}(L_1,L_2)+\pi_{R_1}(L_2)(1-\pi_{R_2}(L_1))g_{L_1}(L_1)\\
&+(1-\pi_{R_1}(L_2))\pi_{R_2}(L_1)g_{L_2}(L_2)+(1-\pi_{R_1}(L_2))(1-\pi_{R_2}(L_1))g_0\}]=0\\
\Leftrightarrow&\mathbb{E}[\pi_{R_1}(L_2)\pi_{R_2}(L_1)g_{L_1,L_2}(L_1,L_2)+\pi_{R_1}(L_2)(1-\pi_{R_2}(L_1))g_{L_1}(L_1)\\
&+(1-\pi_{R_1}(L_2))\pi_{R_2}(L_1)g_{L_2}(L_2)+(1-\pi_{R_1}(L_2))(1-\pi_{R_2}(L_1))g_0|L_1,L_2]=0
\end{aligned} \quad \text{(B.6)}$$

Additionally, we also need the functions in $\Gamma^{obs,}$ to have expectation zero by the definition of influence function:

$$\begin{aligned}
&\mathbb{E}[\pi_{R_1}(L_2)\pi_{R_2}(L_1)g_{L_1,L_2}(L_1,L_2)+\pi_{R_1}(L_2)(1-\pi_{R_2}(L_1))g_{L_1}(L_1)+\\
&(1-\pi_{R_1}(L_2))\pi_{R_2}(L_1)g_{L_2}(L_2)+(1-\pi_{R_1}(L_2))(1-\pi_{R_2}(L_1))g_0]=0
\end{aligned} \quad \text{(B.7)}$$

After simplifying the above constraints, we are able to obtain the following simplified constraints which reveals that the degree of freedom of Block Parallel Missing Data graphical model is one.

APPENDIX B. DERIVATION OF EFFICIENT INFLUENCE FUNCTION IN TWO-VARIABLES MODEL

$$g_{L_2}(L_2) = -\{\mathbb{E}[\pi_{R_2}(L_1)|L_2]\}^{-1}\{1 - \mathbb{E}[\pi_{R_2}(L_1)|L_2]\} \cdot g_0 \tag{B.8}$$

$$g_{L_1}(L_1) = -\{\mathbb{E}[\pi_{R_1}(L_2)|L_1]\}^{-1}\{1 - \mathbb{E}[\pi_{R_1}(L_2)|L_1]\} \cdot g_0 \tag{B.9}$$

$$\begin{aligned} g_{L_1,L_2}(L_1, L_2) = \{&\frac{(1 - \pi_{R_2}(L_1))(1 - \mathbb{E}[\pi_{R_1}(L_2)|L_1])}{\pi_{R_2}(L_1)\mathbb{E}[\pi_{R_1}(L_2)|L_1]} + \frac{(1 - \pi_{R_1}(L_2))(1 - \mathbb{E}[\pi_{R_2}(L_1)|L_2])}{\pi_{R_1}(L_2)\mathbb{E}[\pi_{R_2}(L_1)|L_2]} \\ &- \frac{(1 - \pi_{R_1}(L_2))(1 - \pi_{R_2}(L_1))}{\pi_{R_1}(L_2)\pi_{R_2}(L_1)}\} \cdot g_0 \end{aligned} \tag{B.10}$$

In other words, as long as we determine the value of g_0, assuming that we also have access to the knowledge of $\pi_{R_1}(L_2)$ and $\pi_{R_2}(L_1)$, we can further determine the function $g(R_1, R_1L_1, R_2, R_2L_2)$.

Proposition 1. *Given IF^* derived in equation (3.4), the efficient observed data influence function EIF^{obs} is*

$$EIF^{obs} = IF^* - \Pi[IF^*|g(R_1, L_1R_1, R_2, L_2R_2)] \tag{B.11}$$

where $\Pi(\cdot)$ denotes the projection operator.

APPENDIX B. DERIVATION OF EFFICIENT INFLUENCE FUNCTION IN TWO-VARIABLES MODEL

Proof.

$$\begin{aligned}\mathbb{E}[gg^T] =&\mathbb{E}\{\frac{\pi_{R_2}(L_1)(1-\pi_{R_1}(L_2))}{\pi_{R_1}(L_2)}\Psi(L_2)^2+\frac{\pi_{R_1}(L_2)(1-\pi_{R_2}(L_1))}{\pi_{R_2}(L_1)}\Psi(L_1)^2\\&+2(1-\pi_{R_1}(L_2))(1-\pi_{R_2}(L_1))\Psi(L_1)\Psi(L_2)\\&-2\frac{(1-\pi_{R_1}(L_2))(1-\pi_{R_2}(L_1))^2}{\pi_{R_2}(L_1)}\Psi(L_1)\\&-2\frac{(1-\pi_{R_2}(L_1))(1-\pi_{R_1}(L_2))^2}{\pi_{R_1}(L_2)}\Psi(L_2)\\&+(1-\pi_{R_1}(L_2))(1-\pi_{R_2}(L_1))[1+\frac{(1-\pi_{R_1}(L_2))(1-\pi_{R_2}(L_1))}{\pi_{R_1}(L_2)\pi_{R_2}(L_1)}]\},\end{aligned} \tag{B.12}$$

where $\Psi(L_1)=\frac{1-\mathbb{E}\{\pi_{R_1}(L_2)|L_1\}}{\mathbb{E}\{\pi_{R_1}(L_2)|L_1\}}$ and $\Psi(L_2)=\frac{1-\mathbb{E}\{\pi_{R_2}(L_1)|L_2\}}{\mathbb{E}\{\pi_{R_2}(L_1)|L_2\}}$. Denote $g_0^*=\mathbb{E}\{IF^*\cdot g(R_1,L_1R_1,R_2,L_2R_2)^T\}\cdot[\mathbb{E}\{g(R_1,L_1R_1,R_2,L_2R_2)\cdot g(R_1,L_1R_1,R_2,L_2R_2)^T\}]^{-1}\cdot g_0^{-1}$ for short-hand notation,

$$\mathbb{E}[IF^*\times g^T]=(i)+(ii)+(iii)$$

APPENDIX B. DERIVATION OF EFFICIENT INFLUENCE FUNCTION IN TWO-VARIABLES MODEL

where

$$(i) = \mathbb{E}[\frac{h(L_1, L_2)}{\pi_{R_1}(L_2)\pi_{R_2}(L_1)}\{\pi_{R_1}(L_2)(1-\pi_{R_2}(L_1))\Psi(L_1) + \pi_{R_2}(L_1)(1-\pi_{R_1}(L_2)\Psi(L_2) - (1-\pi_{R_1}(L_2))(1-\pi_{R_2}(L_1))\}],$$

$$(ii) = \mathbb{E}[\frac{h(L_1, L_2)(1-\pi_{R_1}(L_2))}{\pi_{R_2}(L_1)}\mathbb{E}\{(1-\pi_{R_2}(L_1))(\Psi(L_1)-1)|L_2\}|R_1=1, R_2=1],$$

$$(iii) = \mathbb{E}[\frac{h(L_1, L_2)(1-\pi_{R_2}(L_1))}{\pi_{R_1}(L_2)}\mathbb{E}\{(1-\pi_{R_1}(L_2))(\Psi(L_2)-1)|L_1\}|R_1=1, R_2=1].$$

Then

$$\begin{aligned} EIF^{obs} = IF^* - g_0^*[R_1R_2\{&\frac{1-\pi_{R_2}(L_1)}{\pi_{R_2}(L_1)\mathbb{E}(\pi_{R_1}(L_2)|L_1)} + \frac{1-\pi_{R_1}(L_2)}{\pi_{R_1}(L_2)\mathbb{E}(\pi_{R_2}(L_1)|L_2)} \\ &- \frac{1-\pi_{R_1}(L_2)\pi_{R_2}(L_1)}{\pi_{R_1}(L_2)\pi_{R_2}(L_1)}\} \\ &- R_1(1-R_2)\frac{1-\mathbb{E}(\pi_{R_1}(L_2)|L_1)}{\mathbb{E}(\pi_{R_1}(L_2)|L_1)} - R_2(1-R_1)\frac{1-\mathbb{E}(\pi_{R_2}(L_1)|L_2)}{\mathbb{E}(\pi_{R_2}(L_1)|L_2)} \\ &+ (1-R_1)(1-R_2)] \end{aligned}$$

It is equivalent to

$$\begin{aligned} EIF^{obs} = IF^* - \mathbb{E}\left\{IF^* \cdot g\left(R_1, L_1R_1, R_2, L_2R_2\right)^T\right\} \cdot \\ \left[\mathbb{E}\left\{g\left(R_1, L_1R_1, R_2, L_2R_2\right) \cdot g\left(R_1, L_1R_1, R_2, L_2R_2\right)^T\right\}\right]^{-1} \cdot g\left(R_1, L_1R_1, R_2, L_2R_2\right) \end{aligned}$$

□

Bibliography

[1] K. Mohan, J. Pearl, and J. Tian, "Graphical models for inference with missing data," in *Advances in neural information processing systems*, 2013, pp. 1277–1285.

[2] K. Mohan and J. Pearl, "Graphical models for recovering probabilistic and causal queries from missing data," in *Advances in Neural Information Processing Systems*, 2014, pp. 1520–1528.

[3] J. Pearl, *Causality: models, reasoning and inference*. Springer, 2000, vol. 29.

[4] K. Mohan and J. Pearl, "Graphical models for processing missing data," *arXiv preprint arXiv:1801.03583*, 2018.

[5] A. Tsiatis, *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[6] D. Malinsky, I. Shpitser, and E. J. T. Tchetgen, "Semiparametric infer-

ence for non-monotone missing-not-at-random data: the no self-censoring model," *arXiv preprint arXiv:1909.01848*, 2019.

Vita

1996 Born, Shanxi, China

2018 B.S., Statistics, Central University of Finance and Economics, Beijing, China

2019 M.S.E., Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, Maryland